Weekly Reading

江昊宇

2022-07-29

ArcFace: Additive Angular Margin Loss for Deep Face Recognition

* Imperial College London

- Motivation
  - Propose an Additive Angular Margin Loss (ArcFace) to obtain highly discriminative features for face recognition
  - ArcFace consistently outperforms the state-of-the-art and can be easily implemented with negligible computational overhead
- Datasets

Datasets	#Identity	#Image/Video
CASIA [43]	10K	0.5M
VGGFace2 [6]	9.1K	3.3M
MS1MV2	85K	5.8M
MS1M-DeepGlint [2]	87K	3.9M
Asian-DeepGlint [2]	94 K	2.83M
LFW [13]	5,749	13,233
CFP-FP [30]	500	7,000
AgeDB-30 [22]	568	16,488
CPLFW [48]	5,749	11,652
CALFW [49]	5,749	12,174
YTF [40]	1,595	3,425
MegaFace [15]	530 (P)	1M (G)
IJB-B [39]	1,845	76.8K
IJB-C [21]	3,531	148.8K
Trillion-Pairs [2]	5,749 (P)	1.58M (G)
iQIYI-VID [20]	4,934	172,835

Table 1. Face datasets for training and testing. "(P)" and "(G)" refer to the probe and gallery set, respectively.

ArcFace: Additive Angular Margin Loss for Deep Face Recognition

* Imperial College London

- Methods
  - Additive Angular Margin Loss

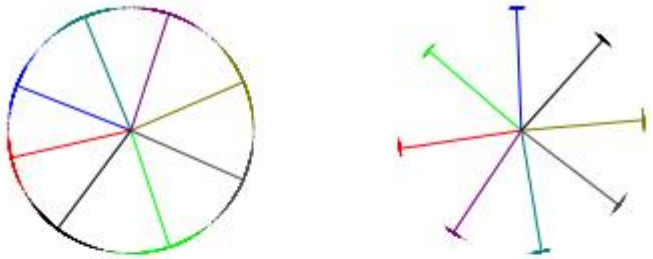
$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}, \tag{1}$$

$$b_j = 0 \qquad W_j^T x_i = \|W_j\|\,\|x_i\|\cos\theta_j$$

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}}. \tag{2}$$

$$L_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}}. \tag{3}$$

  - Comparison with SphereFace and CosFace



(a) Softmax (b) ArcFace

Figure 3. Toy examples under the softmax and ArcFace loss on 8 identities with 2D features. Dots indicate samples and lines refer to the centre direction of each identity. Based on the feature normalisation, all face features are pushed to the arc space with a fixed radius. The geodesic distance gap between closest classes becomes evident as the additive angular margin penalty is incorporated.

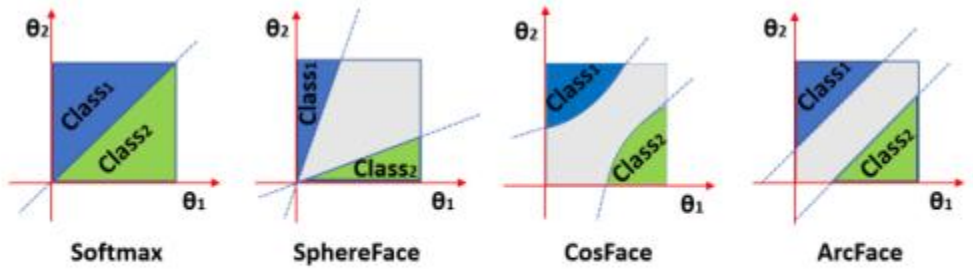


Figure 5. Decision margins of different loss functions under binary classification case. The dashed line represents the decision boundary, and the grey areas are the decision margins.

ArcFace: Additive Angular Margin Loss for Deep Face Recognition

* Imperial College London

- Methods

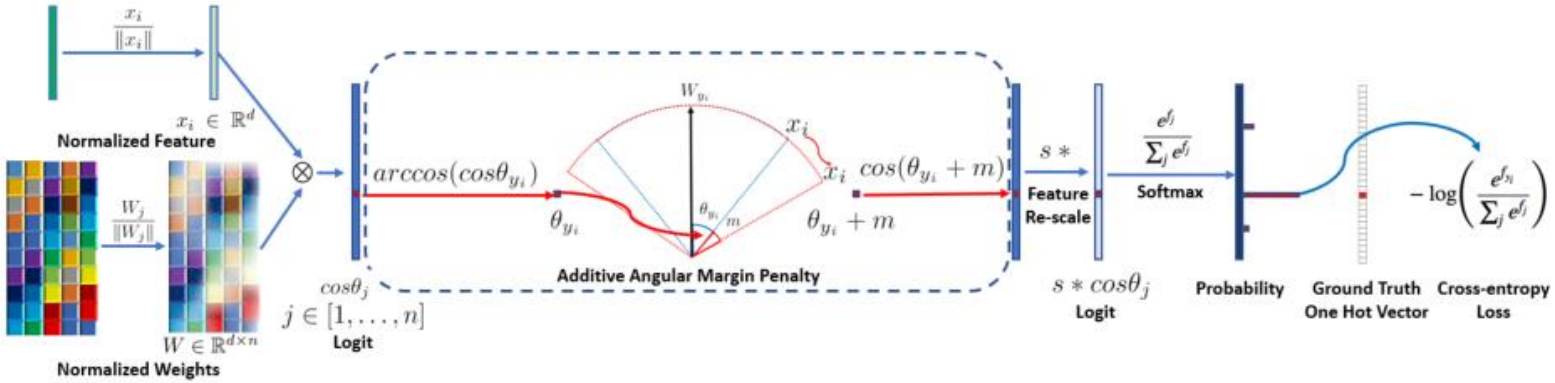


Figure 2. Training a DCNN for face recognition supervised by the ArcFace loss. Based on the feature x_i and weight W normalisation, we get the $\cos\theta_j$ (logit) for each class as $W_j^T x_i$. We calculate the $arccos\theta_{y_i}$ and get the angle between the feature x_i and the ground truth weight W_{y_i}. In fact, W_j provides a kind of centre for each class. Then, we add an angular margin penalty m on the target (ground truth) angle θ_{y_i}. After that, we calculate $\cos(\theta_{y_i} + m)$ and multiply all logits by the feature scale s. The logits then go through the softmax function and contribute to the cross entropy loss.

Algorithm 1 The Pseudo-code of ArcFace on MxNet

Input: Feature Scale s, Margin Parameter m in Eq. 3, Class Number n, Ground-Truth ID gt.

1. x = mx.symbol.L2Normalization (x, mode = 'instance')
2. W = mx.symbol.L2Normalization (W, mode = 'instance')
3. fc7 = mx.sym.FullyConnected (data = x, weight = W, no_bias = True, num_hidden = n)
4. original_target_logit = mx.sym.pick (fc7, gt, axis = 1)
5. theta = mx.sym.arccos (original_target_logit)
6. marginal_target_logit = mx.sym.cos (theta + m)
7. one_hot = mx.sym.one_hot (gt, depth = n, on_value = 1.0, off_value = 0.0)
8. fc7 = fc7 + mx.sym.broadcast_mul (one_hot, mx.sym.expand_dims (marginal_target_logit - original_target_logit, 1))
9. fc7 = fc7 * s

Output: Class-wise affinity score $fc7$.

ArcFace: Additive Angular Margin Loss for Deep Face Recognition
* Imperial College London

- Methods
  - Other Losses

By combining all of the margin penalties, we implement SphereFace, ArcFace and CosFace in an united framework with m_1, m_2 and m_3 as the hyper-parameters.

$$L_4 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(m_1\theta_{y_i}+m_2)-m_3)}}{e^{s(\cos(m_1\theta_{y_i}+m_2)-m_3)}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}} \quad (4)$$

  - Intra-Loss

$$L_5 = L_2 + \frac{1}{\pi N}\sum_{i=1}^{N}\theta_{y_i}. \quad (5)$$

  - Inter-Loss

$$L_6 = L_2 - \frac{1}{\pi N\,(n-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq y_i}^{n}\arccos(W_{y_i}^{T}W_j). \quad (6)$$

  - Triplet-loss

$$\arccos(x_i^{pos}x_i) + m \leq \arccos(x_i^{neg}x_i).$$

ArcFace: Additive Angular Margin Loss for Deep Face Recognition

* Imperial College London

- Results

Loss Functions	LFW	CFP-FP	AgeDB-30
ArcFace (0.4)	99.53	95.41	94.98
ArcFace (0.45)	99.46	95.47	94.93
ArcFace (0.5)	**99.53**	**95.56**	**95.15**
ArcFace (0.55)	99.41	95.32	95.05
SphereFace [18]	99.42	-	-
SphereFace (1.35)	99.11	94.38	91.70
CosFace [37]	99.33	-	-
CosFace (0.35)	99.51	95.44	94.56
CM1 (1, 0.3, 0.2)	99.48	95.12	94.38
CM2 (0.9, 0.4, 0.15)	99.50	95.24	94.86
Softmax	99.08	94.39	92.33
Norm-Softmax (NS)	98.56	89.79	88.72
NS+Intra	98.75	93.81	90.92
NS+Inter	98.68	90.67	89.50
NS+Intra+Inter	98.73	94.00	91.41
Triplet (0.35)	98.98	91.90	89.98
ArcFace+Intra	99.45	95.37	94.73
ArcFace+Inter	99.43	95.25	94.55
ArcFace+Intra+Inter	99.43	95.42	95.10
ArcFace+Triplet	99.50	95.51	94.40

Table 2. Verification results (%) of different loss functions ([CASIA, ResNet50, loss*]).

	NS	ArcFace	IntraL	InterL	TripletL
W-EC	44.26	14.29	8.83	46.85	-
W-Inter	69.66	71.61	31.34	75.66	-
Intra1	50.50	38.45	17.50	52.74	41.19
Inter1	59.23	65.83	24.07	62.40	50.23
Intra2	33.97	28.05	12.94	35.38	27.42
Inter2	65.60	66.55	26.28	67.90	55.94

Table 3. The angle statistics under different losses ([CASIA, ResNet50, loss*]). Each column denotes one particular loss. "W-EC" refers to the mean of angles between W_j and the corresponding embedding feature centre. "W-Inter" refers to the mean of minimum angles between W_j's. "Intra1" and "Intra2" refer to the mean of angles between x_i and the embedding feature centre on CASIA and LFW, respectively. "Inter1" and "Inter2" refer to the mean of minimum angles between embedding feature centres on CASIA and LFW, respectively.

Method	#Image	LFW	YTF
DeepID [32]	0.2M	99.47	93.20
Deep Face [33]	4.4M	97.35	91.4
VGG Face [24]	2.6M	98.95	97.30
FaceNet [29]	200M	99.63	95.10
Baidu [16]	1.3M	99.13	-
Center Loss [38]	0.7M	99.28	94.9
Range Loss [46]	5M	99.52	93.70
Marginal Loss [9]	3.8M	99.48	95.98
SphereFace [18]	0.5M	99.42	95.0
SphereFace+ [17]	0.5M	99.47	-
CosFace [37]	5M	99.73	97.6
MS1MV2, R100, ArcFace	5.8M	**99.83**	**98.02**

Table 4. Verification performance (%) of different methods on LFW and YTF.

Method	LFW	CALFW	CPLFW
HUMAN-Individual	97.27	82.32	81.21
HUMAN-Fusion	99.85	86.50	85.24
Center Loss [38]	98.75	85.48	77.48
SphereFace [18]	99.27	90.30	81.40
VGGFace2 [6]	99.43	90.57	84.00
MS1MV2, R100, ArcFace	**99.82**	**95.45**	**92.08**

Table 5. Verification performance (%) of open-sourced face recognition models on LFW, CALFW and CPLFW.

Methods	Id (%)	Ver (%)
Softmax [18]	54.85	65.92
Contrastive Loss[18, 32]	65.21	78.86
Triplet [18, 29]	64.79	78.32
Center Loss[38]	65.49	80.14
SphereFace [18]	72.729	85.561
CosFace [37]	77.11	89.88
AM-Softmax [35]	72.47	84.44
SphereFace+ [17]	73.03	-
CASIA, R50, ArcFace	77.50	92.34
CASIA, R50, ArcFace, R	91.75	93.69
FaceNet [29]	70.49	86.47
CosFace [37]	82.72	96.65
MS1MV2, R100, ArcFace	81.03	96.98
MS1MV2, R100, CosFace	80.56	96.56
MS1MV2, R100, ArcFace, R	98.35	98.48
MS1MV2, R100, CosFace, R	97.91	97.91

Table 6. Face identification and verification evaluation of different methods on MegaFace Challenge1 using FaceScrub as the probe set. "Id" refers to the rank-1 face identification accuracy with 1M distractors, and "Ver" refers to the face verification TAR at 10^{-6} FAR. "R" refers to data refinement on both probe set and 1M distractors. ArcFace obtains state-of-the-art performance under both small and large protocols.

ArcFace: Additive Angular Margin Loss for Deep Face Recognition

* Imperial College London

- Results

Method	IJB-B	IJB-C
ResNet50 [6]	0.784	0.825
SENet50 [6]	0.800	0.840
ResNet50+SENet50 [6]	0.800	0.841
MN-v [42]	0.818	0.852
MN-vc [42]	0.831	0.862
ResNet50+DCN(Kpts) [41]	0.850	0.867
ResNet50+DCN(Divs) [41]	0.841	0.880
SENet50+DCN(Kpts) [41]	0.846	0.874
SENet50+DCN(Divs) [41]	0.849	0.885
VGG2, R50, ArcFace	0.898	0.921
MS1MV2, R100, ArcFace	**0.942**	**0.956**

Table 7. 1:1 verification TAR (@FAR=1e-4) on the IJB-B and IJB-C dataset.

Method	Id (@FPR=1e-3)	Ver(@FPR=1e-9)
CASIA	26.643	21.452
MS1MV2	80.968	78.600
DeepGlint-Face	80.331	78.586
MS1MV2+Asian	**84.840** (1st)	80.540
CIGIT_IRSEC	84.234 (2nd)	**81.558** (1st)

Table 8. Identification and verification results (%) on the Trillion-Pairs dataset. ([Dataset*, ResNet100, ArcFace])

Method	MAP(%)
MS1MV2+Asian, R100, ArcFace	79.80
+ MLP	86.40
+ Ensemble	88.26
+ Context	**88.65 (1st)**
Other Participant	87.66 (2nd)

Table 9. MAP of our method on the iQIYI-VID test set. "MLP" refers to a three-layer fully connected network trained on the iQIYI-VID training data.

iQIYI-VID: A Large Dataset for Multi-modal Person Identification

- Motivation
  - Traditional research, such as face recognition, person re-identification, and speaker recognition, often focuses on a single modal of information, which is inadequate to handle all the situations in practice
  - Multimodal person identification is a more promising way that we can jointly utilize face, head, body, audio features, and so on
- Pipeline

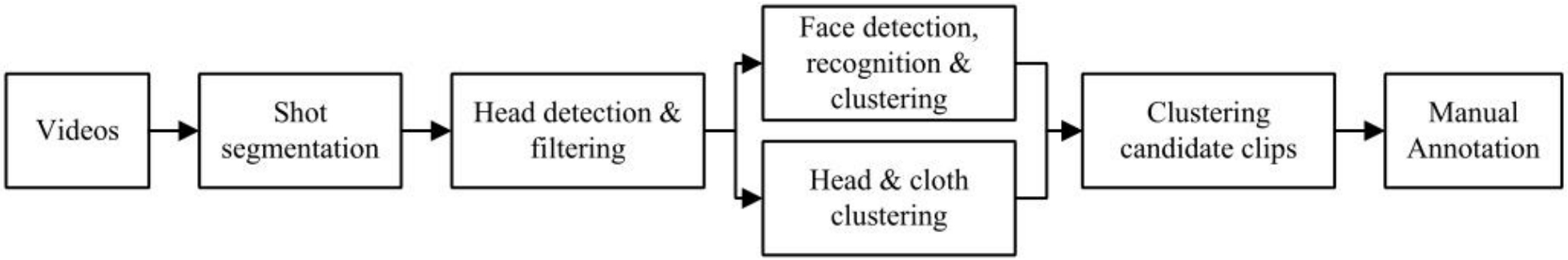


Figure 1: The process of building the iQIYI-VID dataset.

- Pipeline
  - Extracting video clips
    - Each raw video is segmented into shots according to the dissimilarity between consecutive frames
  - Automatic filtering by head detection.
    - A valid frame is defined as a frame in which only one head is detected, or the biggest head is three times larger than the other heads. A valid clip is defined as a clip whose valid frames exceed a ratio of 30%.
  - Obtaining candidate clips for each identity
    - We cluster the clips from the same video by the faces and clothes information. The face and clothes in each frame are paired according to their relative position, and the identities of faces are propagated to the clothes with the face-clothes pairs. After that, each clothes cluster can get an ID through majority voting.
  - Final manual filtering
    - All clusters of clips are cleaned by a manual annotation process. The manual labeling was repeated twice by different labelers to ensure a high quality.

iQIYI-VID: A Large Dataset for Multi-modal Person Identification

- Details
  - It is composed of 600K video clips of 5,000 celebrities.These video clips are extracted from 400K hours of online videos of various types, ranging from movies, variety shows, TV series, to news broadcasting

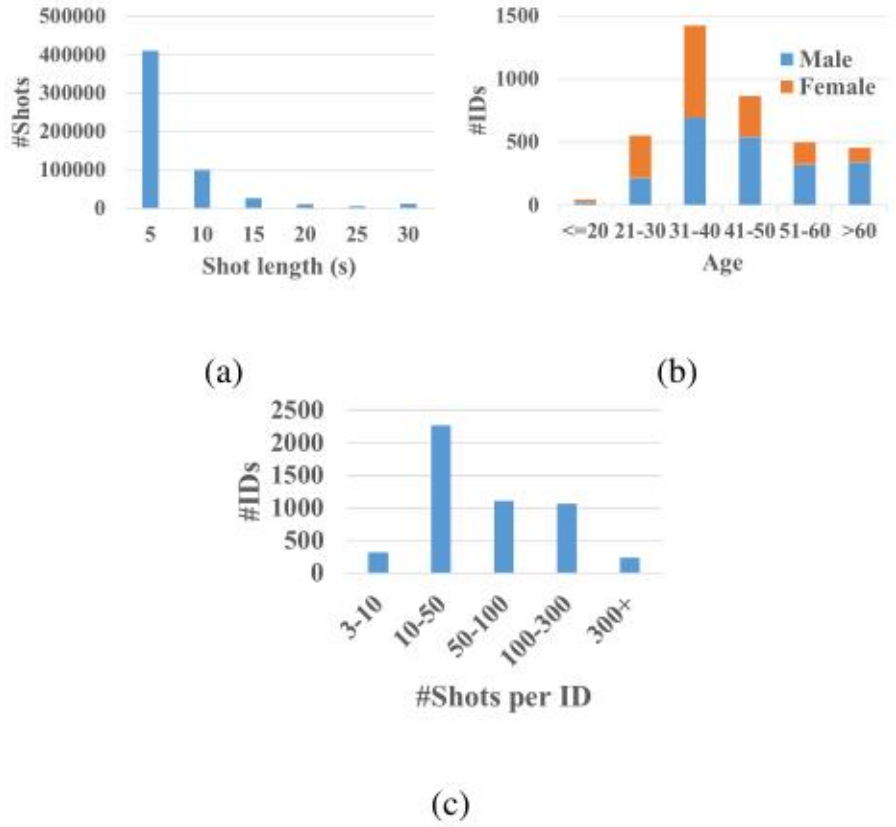


Figure 2: Data distributions.

iQIYI-VID: A Large Dataset for Multi-modal Person Identification

- Method

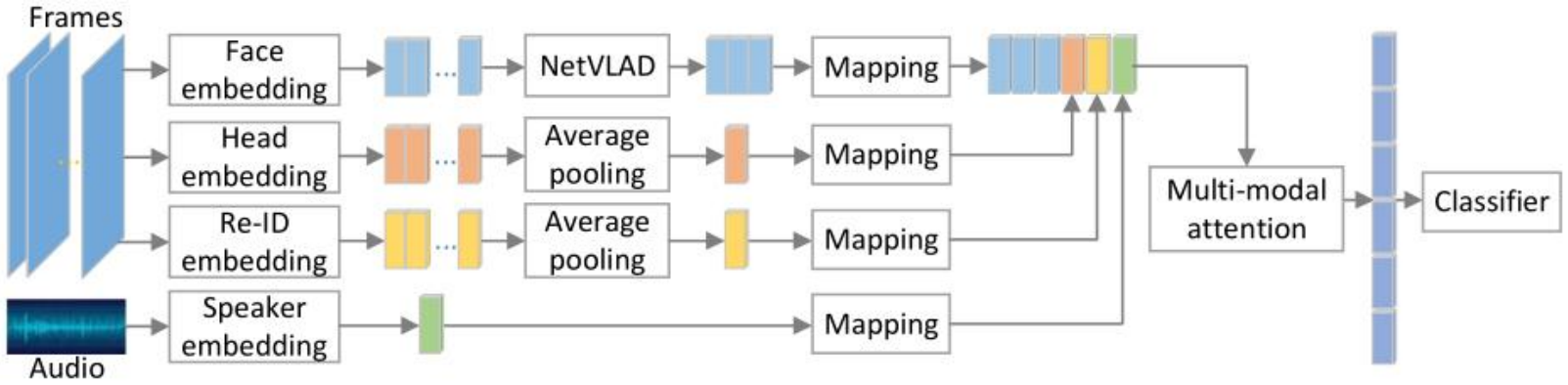


Figure 3: The flowchart of our multi-modal person identification method. It begins with extracting raw features of face, head, body, and audio. The raw features of face, head, and body are transformed into video-level features by an adapted NetVLAD module or Average Pooling. Then they are transformed to the same length by feature mapping. Different modal of features are combined by multi-modal attention, and then fed to a classifier.

- Face
  - SSH、ArcFace
- Head
  - YOLO V2、ArcFace
- Body
  - SSH、Alignedreid++
- Audio
  - ResNet34

- Method
  - Multi-modal Attention module (MMA)

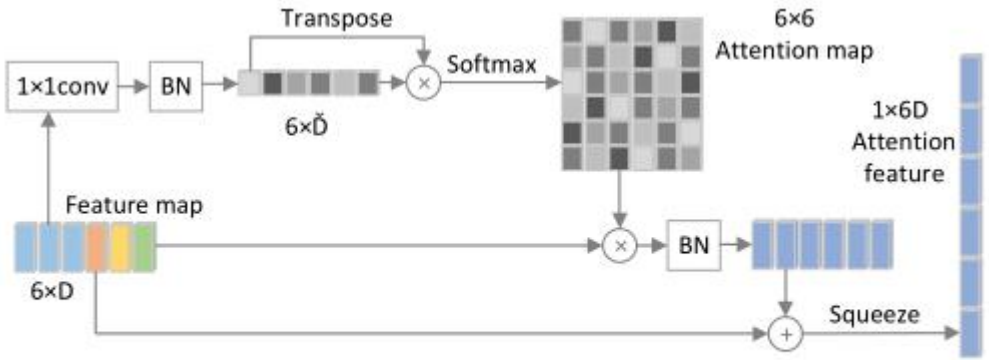


Figure 5: Multi-modal attention (MMA) module.

We propose a Multi-modal Attention module (MMA) to fuse different modal of features together, as shown in Figure 5. The input of MMA is a feature map $X \in \mathbb{R}^{D\times 6}$, where 3 of the 6 features come from face, and the other 3 features come from head, body, and audio, as shown in Figure 3. Each feature has the length of D. The feature map X is transformed into a feature space $F \in \mathbb{R}^{\check{D}\times 6}$ by $W_F \in \mathbb{R}^{\check{D}\times D}$ to calculate the attention $Y \in \mathbb{R}^{6\times 6}$, where

$$Y_{i,j} = \frac{\exp(Z_{i,j})}{\sum_{i=1}^{6}\exp(Z_{i,j})}, \tag{1}$$

with

$$Z = F^{\top}F,\ F = W_F X. \tag{2}$$

Here Z is the Gram matrix of feature map F, which captures the feature correlation [13] that is widely used for image style transfer [14].

The fused feature map $O \in \mathbb{R}^{D\times 6}$ is obtained by

$$O = X + \gamma XY, \tag{3}$$

where γ is a scalar parameter to control the strength of attention. At last the fused feature map is squeezed into a vector and fed to the classifier.

iQIYI-VID: A Large Dataset for Multi-modal Person Identification

- Results

Table 2: Comparison to the state-of-art.

Modal	MAP (%)
ArcFace [11]	88.65
He *et al.* [2]	89.00
Ours	89.70

Table 3: Results of different modal of features and modules.

Modal	MAP (%)
Face	85.19
Head	54.32
Audio	11.79
Body	5.14
Face+Head	87.16
Face+Head+Audio	87.69
Face+Head+Audio+Body	87.80
Ensemble	89.24
+NetVLAD	89.46
+NetVLAD+MMA	89.70

iQIYI-VID: A Large Dataset for Multi-modal Person Identification

- Cases

Figure 6: Challenging cases for head recognition. The hair style and accessories of the same actor changed dramatically.

(a) (b)

Figure 7: Challenging cases for body recognition. (a) An actress changes style in different episodes. (b) Different actors dress the same uniform.

Figure 8: An example of video retrieval results. When adding more and more features inside, the results get much better than face recognition alone. The number above each image indicates the rank of the image within the retrieval results. Positive examples are marked by green boxes, while negative examples are marked in red.

Figure 9: Challenging cases for face recognition. From left to right: profile, occlusion, blur, unusual illumination, and small face.

Thanks